LOFAR Data Services: Now & Next

R. F. Pizzo
Head Science Data Centre Operations

18 September 2024

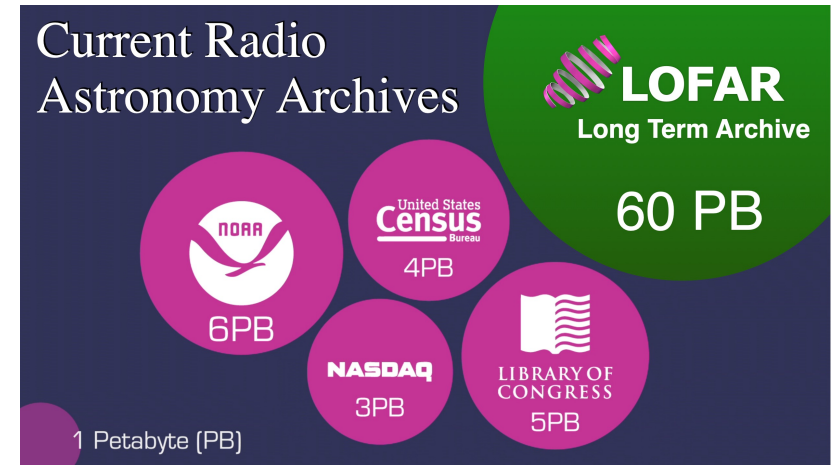# Outline

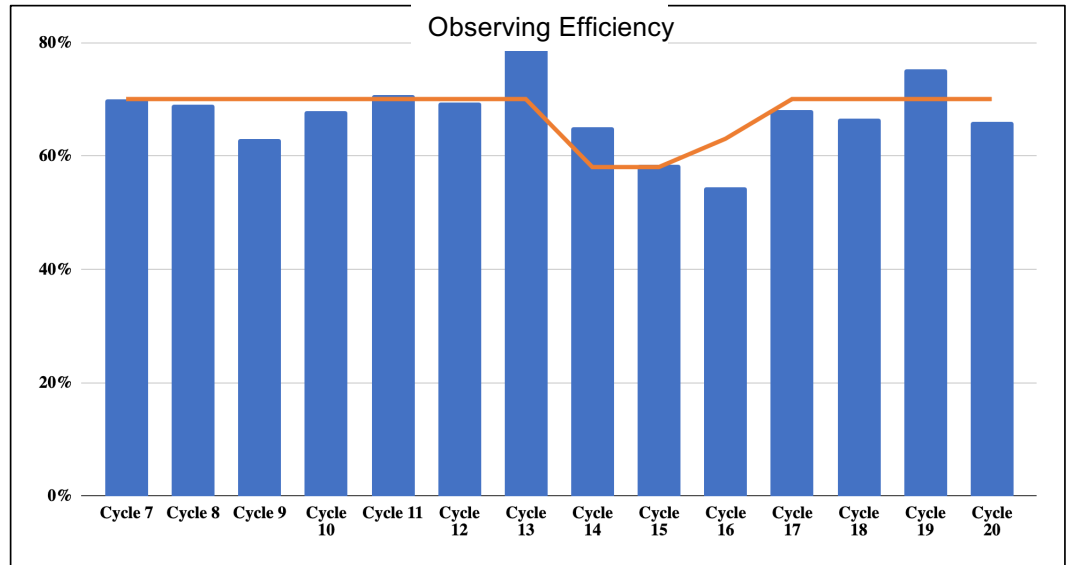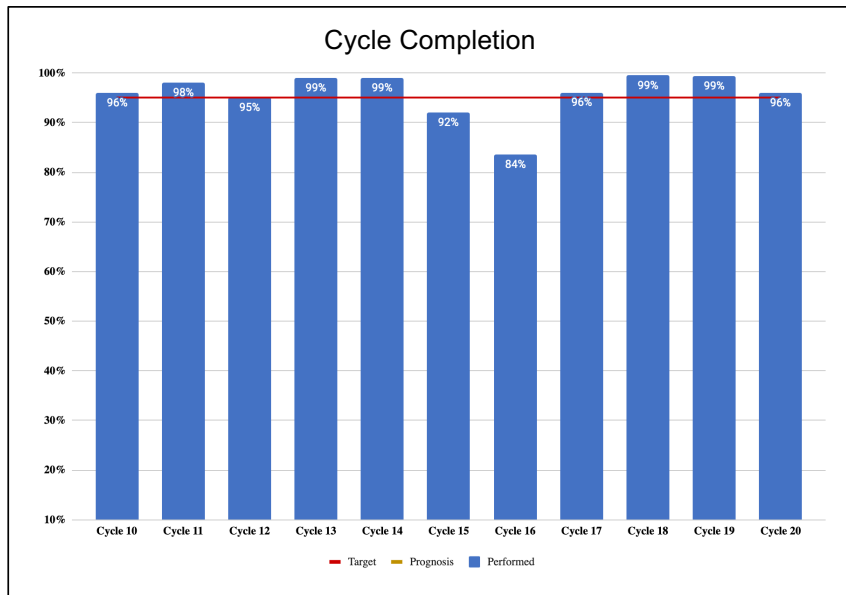- ➢ Current status & successes
- ➢ Future ambition
- ➢ Challenges
- ➢ Resource challenges
- ➢ Data policies

Achievements

➢ Completed 21 operational Cycles

➢ ~65000 hours successfully observed – >70% operational efficiency

➢ Operating a massive array growing in size and capabilities

➢ 60 PB (!) in the LTA- Largest astronomical data collection to date.

➢ Started LTA operations

➢ Supported an ever-growing community

➢ Brought the instrument closer to our users:

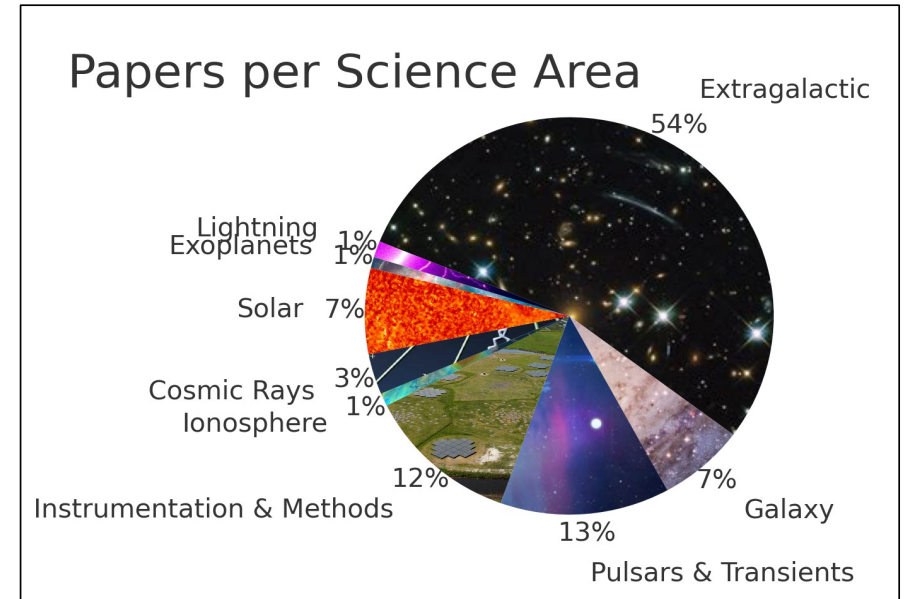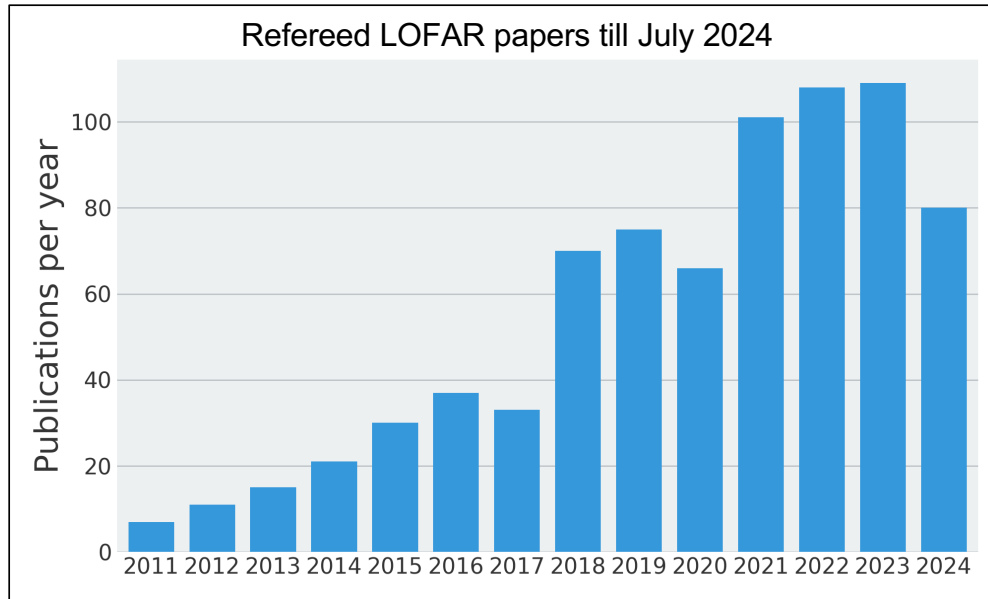 ▪ LOFAR Schools (400+ participants)
  ▪ 60 Busy Weeks
  ▪ Traineeships

Current Radio Astronomy Archives

LOFAR Long Term Archive — 60 PB

NOAA 6PB

United States Census Bureau 4PB

NASDAQ 3PB

LIBRARY OF CONGRESS 5PB

1 Petabyte (PB)

LOFAR School 2024

Observatory Performance

Cycle Completion — Performed percentages by cycle: Cycle 10: 96%, Cycle 11: 98%, Cycle 12: 95%, Cycle 13: 99%, Cycle 14: 99%, Cycle 15: 92%, Cycle 16: 84%, Cycle 17: 96%, Cycle 18: 99%, Cycle 19: 99%, Cycle 20: 96%. (Target, Prognosis, Performed)

Observing Efficiency — Cycles 7 through 20

➤ Target efficiency: 70%

➤ Target completion: 95%

➤ Cycle 20: 96% completed

Lofar Science Output

Refereed LOFAR papers till July 2024

Papers per Science Area

- Extragalactic 54%
- Lightning 1%
- Exoplanets 1%
- Solar 7%
- Cosmic Rays 3%
- Ionosphere 1%
- Instrumentation & Methods 12%
- Pulsars & Transients 13%
- Galaxy 7%

➤ > 750 refereed publications

➤ Publication rate: 2 papers per week – top 10% of all astronomical facilities

5

The LOFAR System: Data Flow…so far

stations: data collection

COBALT: correlation

CEP4 compute cluster

Intermediate data generated through Observatory pipelines

ingest into Long-term archive (3 sites: NL, D, PL). dCache system.

portal
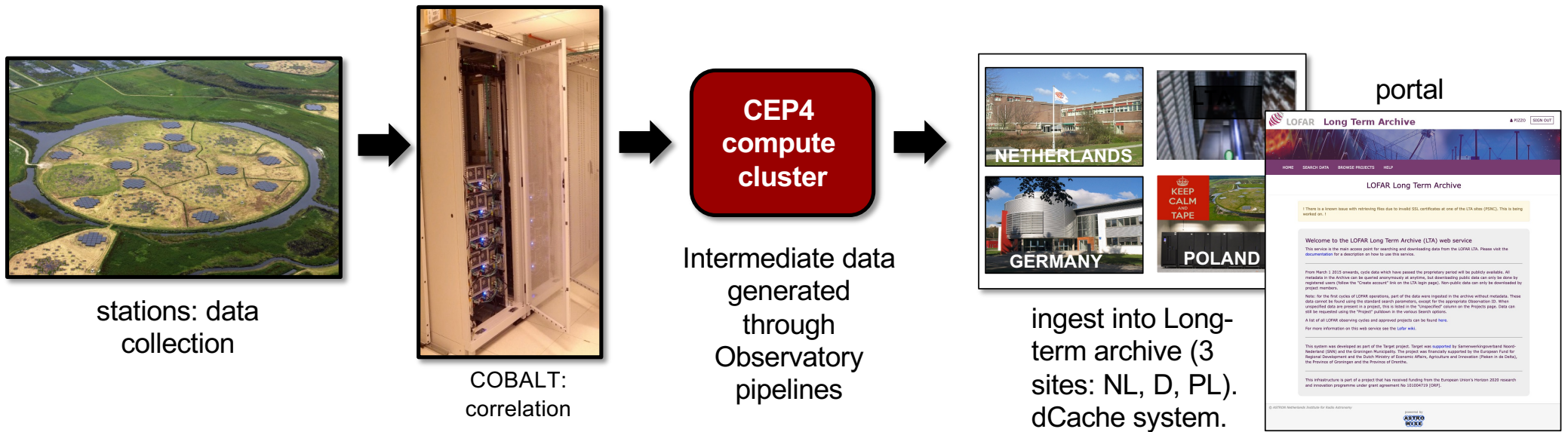
> ➤ Transport, processing and storage of large amounts of data :

- Data flow from all antennas combined: 1.7 Tbyte/s
- To COBALT from station after beamforming: 28 Gbyte/s
- Correlator output to disk: between 2-10 Gbyte/s
- Data storage challenges: ~ 80 TB/h
- Data transfer to the archive: ~10 TB/h
- Archive now: ~ 60 PB in mixed state of reduction and science readiness

CENTRALIZED PROCESSING OF LOFAR DATA...SO FAR

➢ Initial processing in Groningen

➢ Single copy of uncalibrated data on one of the three archive sites

➢ Advanced processing handled by PI's on external facilities (or at the LTA)

➢ Some advanced products released through data releases

Challenges: Data Storage, Access, Distribution & Curation

➢ Storing data gets costly very quickly, especially online storage:
  - ➢ disk- Online-Pb-Year ~ € 100,000
  - ➢ tape - Nearline-Pb-Year ~ € 14,000
  - ➢ one enters a regime where re-observing is cheaper

➢ Data access
  - • Need good interfaces and functionalities that help users to mine the archive and find the data
  - • Intermediate data should be easily retrievable:
    - • user data access limited by dCache overall capacity and bandwidth
    - • LOFAR software distribution required

➢ Curation:
  - • Making available advanced products from users
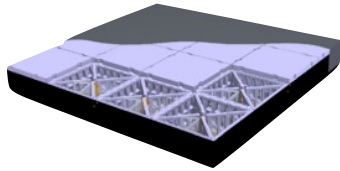   - • Need 'user ingest' and a LOFAR data 'hub'

➢ Distribution
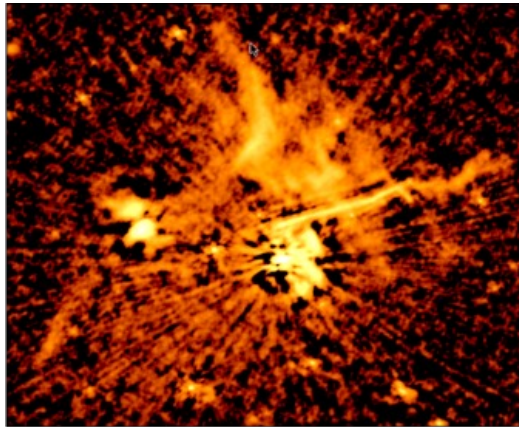 - • Moving large amounts of data is impractical
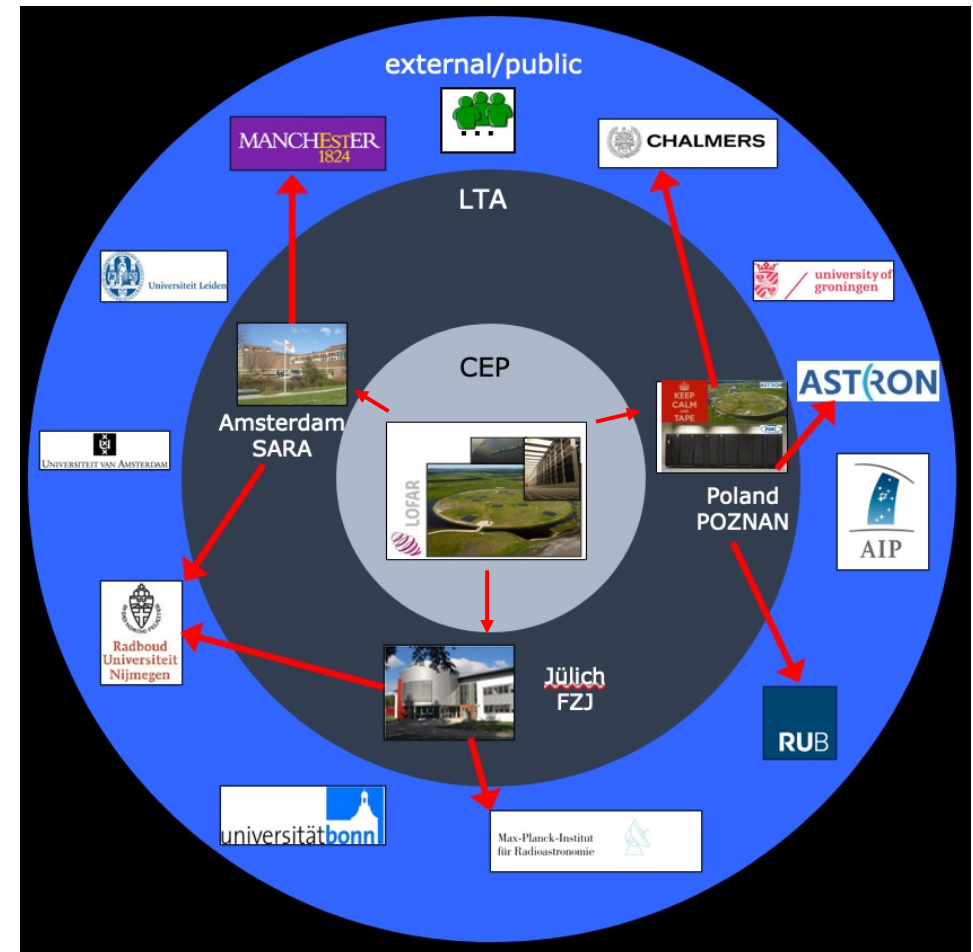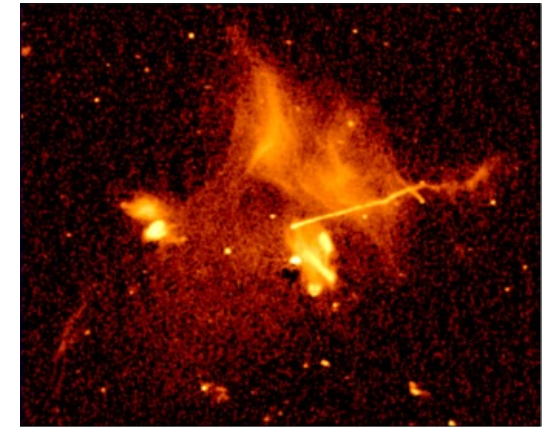
COMING NEXT: LOFAR2.0

High-Band

Low-Band

2x

Scientifically limited

No ionospheric correction

Rich in science

Breakthrough techniques
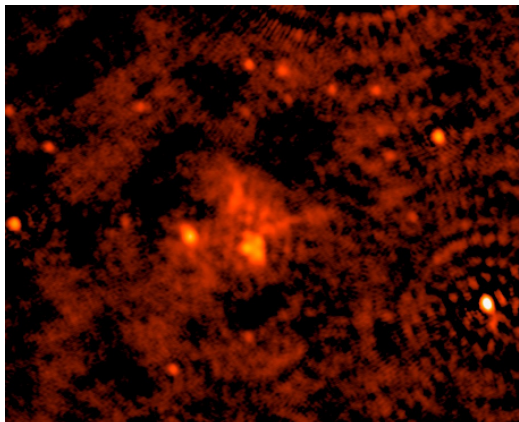
Ionosphere well modeled

LOFAR2.0

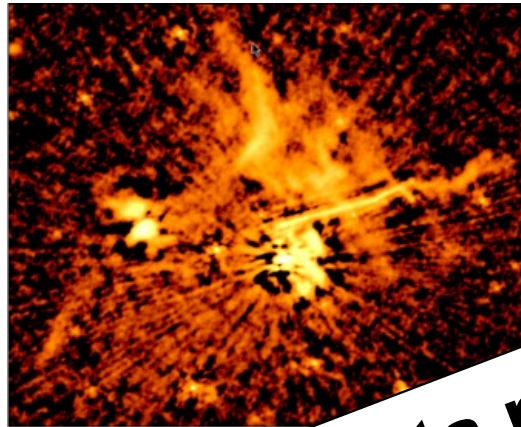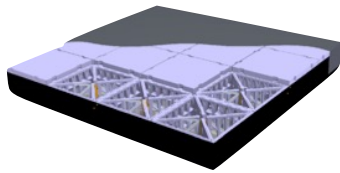AUTOMATIC

The Goal

ASTRON
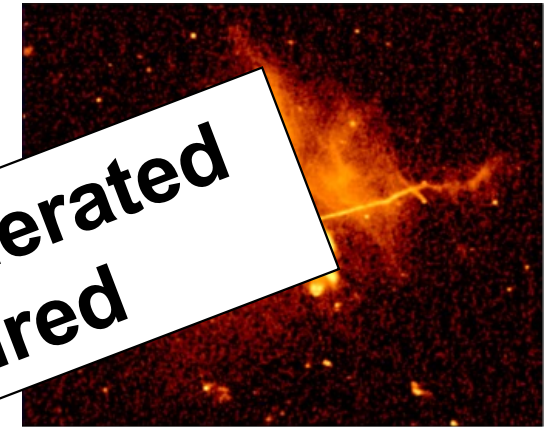Netherlands Institute for Radio Astronomy

COMING NEXT: LOFAR2.0

Scientifically limited

Rich in science

High-Band
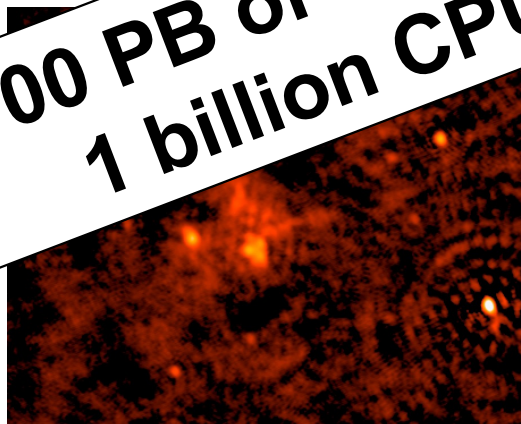
Breakthrough techniques

No ionosph...

Ionosphere well modeled

Low-Band

2x

LOFAR2.0

AUTOMATIC

The Goal

>100 PB of data products generated 1 billion CPU hours required
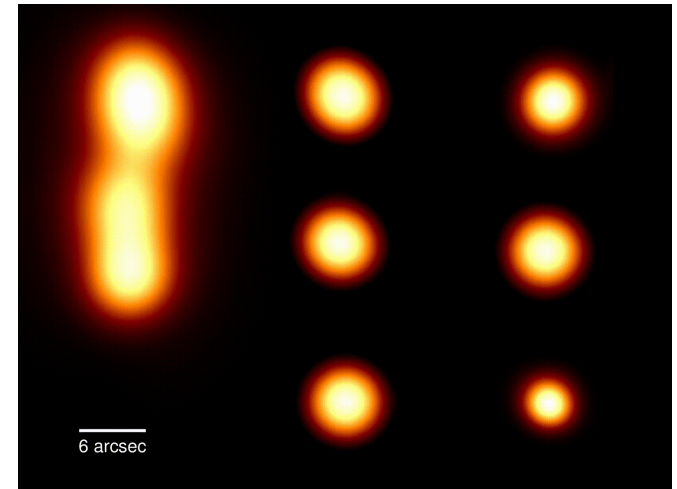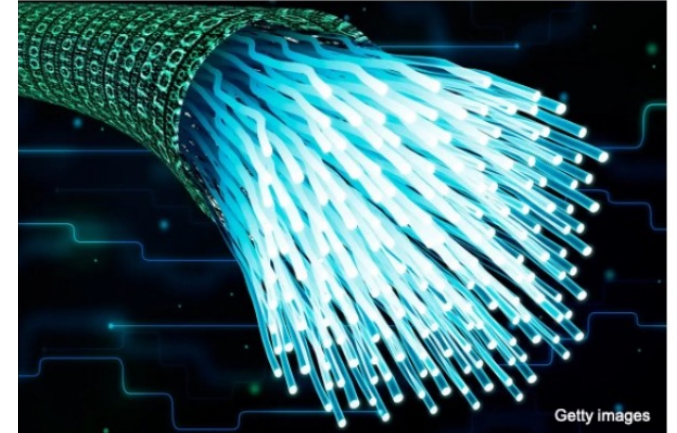
ASTRON
Netherlands Institute for Radio Astronomy

BEYOND LOFAR2.0

- LENSS – LOFAR Enhanced Network for Sharp Surveys
- Upgrade the network (10 → 100Gb/s) for full-FOV, full-res imaging
- Will require high-throughput data processing system deploying innovative algorithms capable of keeping up with the data streaming from the telescope
- Data products generated: 50 PB/year
- Computing: 1 billion CPU hours

Getty images

6 arcsec

Tackling the Challenge: Generating Science-Ready Data

data quality flag

LTA FUTURE

Courtesy SKSP

LTA FUTURE

Courtesy SKSP

- ➤ New operational area compared to LOFAR1

- ➤ LDV:
  - ➤ Reduce data volume at the LTA to reduce operational costs
  - ➤ Streamline data processing operations at the LTA
  - ➤ Prepare ASTRON for LOFAR2 Large Programs

- ➤ LDV *Operations* started early 2023
  - ➤ Current focus: BF data processing (re-quantization) – IF data compression following this year. Savings so far: >5PB + introduction of data quality flag

- ➤ Further valorisation of portions of the archive through data previews (*marriage* with LOFAR2.0 commissioning)

LDV timeline

12
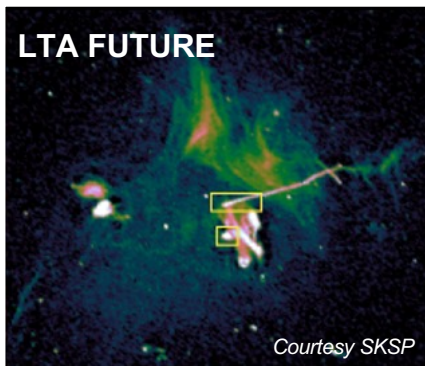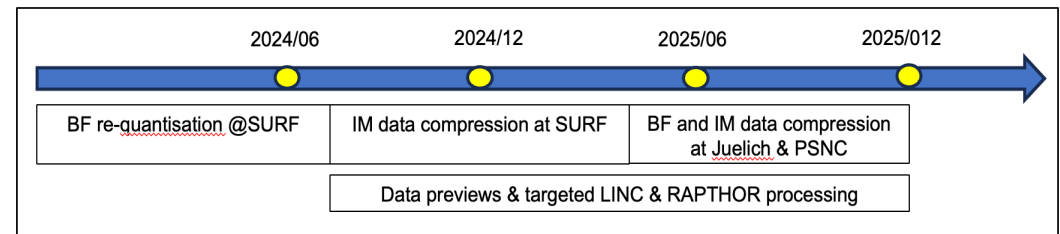
Tackling the Challenges: Forthcoming Tools and Functionalities Developed by the SDC

Support for advanced products, become a 'hub' for access to LOFAR data

For generation of science-ready data
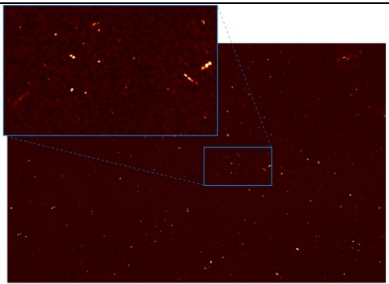
Robust and reliable access to data, use of VO interfaces to publish data, FAIRness

Image courtesy J. Swinbank

Preparing for LOFAR2.0: Data Life Cycle & Early Cycle Data Retirement

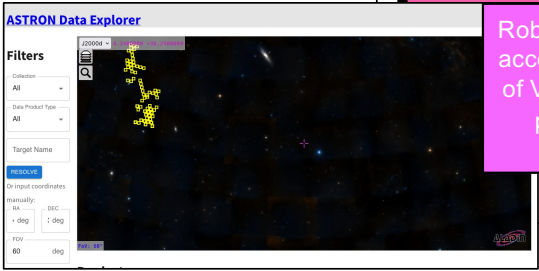| Product type	Example	Retention period
Raw	unprocessed vis.	Not retained
Instrumental	Flagged, compressed vis.	18 months
Intermediate	Direction-independent vis	18 months
Advanced	Images, cubes	Indefinite
Special cases	Unique observations	For discussion

➢ LOFAR2 will generate considerably more data than LOFAR1: ~70 PB intermediate + ~30 PB advanced

➢ Data challenge outstrips current affordable solutions

➢ ILT-board approved a data life cycle:
  • Advanced data products (images, cubes, catalogues) kept indefinitely
  • Intermediate data products will be retired after a period (~18 months), based on available resources
  • Exceptions to be considered in exceptional cases

➢ To prepare for LOFAR2, a first step is taken now: retirement of early LOFAR Cycle data (Cycle 0 till Cycle 6)
  • Timeline: end-2024

The SKA Challenge: Later This Decade
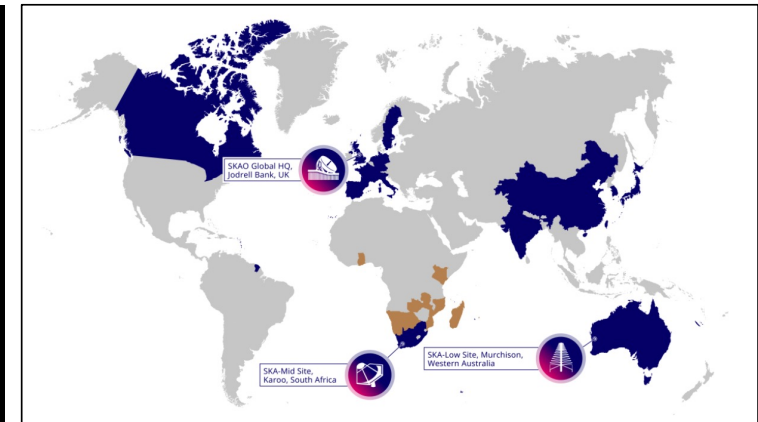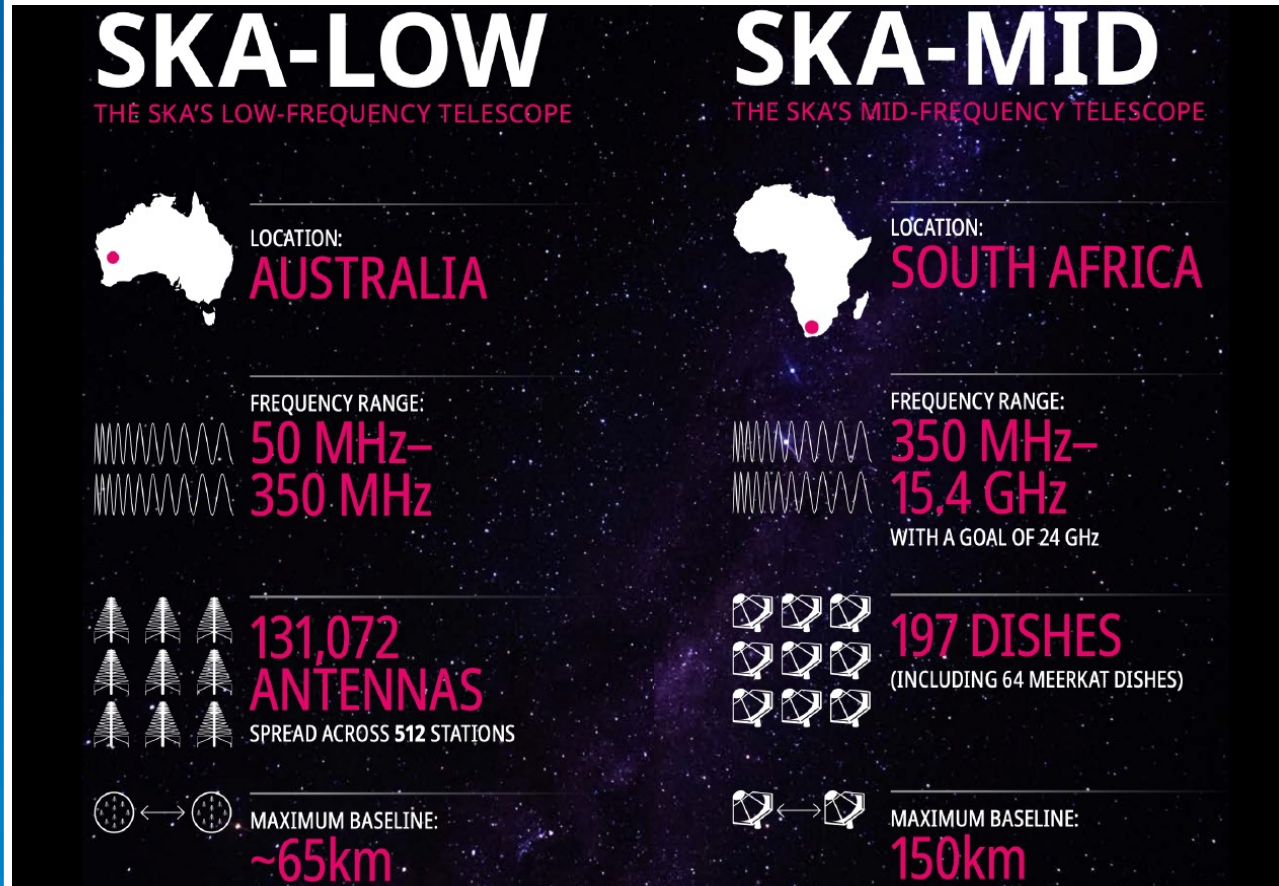
- Science-ready data generated by the SKA observatory: 600 PB/year – new magnitude for astronomical context
- Providing this scientific data repository represents a technical challenge for discovery , analysis and exploitation tasks
- Should include big data lakes and change of paradigm for data access
- SKAO data products provided to a network of SKA regional centers (SRC's) responsible for archiving & data curation, making the data available to the users, enabling scientific discovery

To Conclude

➢ LOFAR is an important technological pathfinder for next-generation data-intensive radio astronomy

➢ We face fundamental challenges related to the data handling and this will become even more acute for the LOFAR upgrades

➢ Demand for putting in place policies that adequately prepare the infrastructure for the future